(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平10－143540

(43)公開日　平成10年(1998)５月29日

(51)Int.Cl.[6]	識別記号	FI		
G06F	17/30	G06F	15/403	340A
H04L	12/54		15/40	310F
	12/58		15/403	350C
		H04L	11/20	101B

審査請求　未請求　請求項の数6　ＯＬ　（全７頁）

(21)出願番号　特願平9－248599

(22)出願日　平成９年(1997)９月12日

(31)優先権主張番号　特願平8－243295

(32)優先日　平８(1996)９月13日

(33)優先権主張国　日本（ＪＰ）

(71)出願人　000003078
株式会社東芝
神奈川県川崎市幸区堀川町72番地

(72)発明者　住田　一男
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(72)発明者　三池　誠司
東京都港区芝浦一丁目１番１号　株式会社東芝本社事務所内

(72)発明者　酒井　哲也
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内

(74)代理人　弁理士　鈴江　武彦　（外６名）

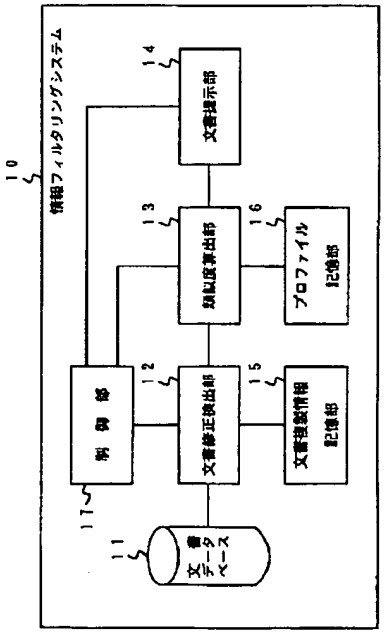
最終頁に続く

(54)【発明の名称】　情報フィルタリング装置および情報フィルタリング方法

(57)【要約】

【課題】非定期的に作成および修正される文書を対象にして、ユーザが必要とする文書のみを絞り込んで提供する情報フィルタリング装置。

【解決手段】文書データベース１１に格納された非定期的に作成および修正される複数の文書は、定期的に読み出され、文書複製情報記憶部１５は、その文書自体またはその文書を情報圧縮したデータを文書複製情報として記憶する。一方、文書修正検出部１２は、前回の読み出し時に作成した文書複製情報と今回読み出した文書情報との違いから、新たに作成または修正された文書を検出する。この検出された文書は、類似度算出部１３によって予めユーザが指定した検索条件との間の類似度が算出される。そして、これらの文書は、類似度算出部１３が算出した類似度に基づき、文書呈示部によってユーザに送信または提示される。



1

【特許請求の範囲】
【請求項１】　非定期的に作成および修正される複数の文書の中から所定の文書を選出してユーザに提示する情報フィルタリング装置において、
新たに作成または修正された文書を検出する文書修正検出手段と、
この文書修正検出手段によって検出された文書と予めユーザが指定した検索条件との間の類似度を算出する類似度算出手段と、
この類似度算出手段により算出された類似度にしたがって前記検出された文書すべてを並べ換え、この並べ換えた順番でこれらの文書をユーザに送信または提示する手段とを具備してなることを特徴とする情報フィルタリング装置。
【請求項２】　前記文書自体または前記文書を情報圧縮したデータを文書複製情報として各文書の文書ＩＤと対応させて格納する文書複製情報記憶手段をさらに具備し、
前記文書修正検出手段は、前記文書複製情報記憶手段に格納された文書複製情報との比較によって前記文書の作成および修正を検出する手段を具備してなることを特徴とする請求項１記載の情報フィルタリング装置。
【請求項３】　非定期的に作成および修正される複数の文書の中から所定の文書を選出してユーザに提示する情報フィルタリング方法において、
新たに作成または修正された文書を検出し、
この検出された文書と予めユーザが指定した検索条件との間の類似度を算出し、
この算出された類似度にしたがって前記検出された文書すべてを並べ換え、
この並べ換えた順番でこれらの文書をユーザに送信または提示することを特徴とする情報フィルタリング方法。
【請求項４】　非定期的に作成および修正される複数の文書の中から所定の文書を選出してユーザに提示する情報フィルタリング方法において、
前記文書自体または前記文書を情報圧縮したデータを文書複製情報として各文書の文書ＩＤと対応させて格納し、
前記格納された文書複製情報との比較によって新たに作成または修正された文書を検出し、
この検出された文書と予めユーザが指定した検索条件との間の類似度を算出し、
この算出された類似度にしたがって前記検出された文書すべてを並べ換え、
この並べ換えた順番でこれらの文書をユーザに送信または提示することを特徴とする情報フィルタリング方法。
【請求項５】　非定期的に作成および修正される複数の文書の中から所定の文書を選出してユーザに提示するためのプログラムであって、
新たに作成または修正された文書を検出し、

2

この検出された文書と予めユーザが指定した検索条件との間の類似度を算出し、
この算出された類似度にしたがって前記検出された文書すべてを並べ換え、
この並べ換えた順番でこれらの文書をユーザに送信または提示するようにコンピュータを動作させるプログラムを記録したコンピュータ読み取り可能な記録媒体。
【請求項６】　非定期的に作成および修正される複数の文書の中から所定の文書を選出してユーザに提示するためのプログラムであって、
前記文書自体または前記文書を情報圧縮したデータを文書複製情報として各文書の文書ＩＤと対応させて格納し、
前記格納された文書複製情報との比較によって新たに作成または修正された文書を検出し、
この検出された文書と予めユーザが指定した検索条件との間の類似度を算出し、
この算出された類似度にしたがって前記検出された文書すべてを並べ換え、
この並べ換えた順番でこれらの文書をユーザに送信または提示するようにコンピュータを動作させるプログラムを記録したコンピュータ読み取り可能な記録媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】この発明は、膨大な数のテキスト記事や文献などの文書からユーザの要求・興味にあったものを選出して定期的にユーザに提供する情報フィルタリング装置および情報フィルタリング方法に関する。
【０００２】
【従来の技術】近年、ワードプロセッサや電子計算機の普及、およびインターネットなどの計算機ネットワークを介した電子メールや電子ニュースの普及に伴ない、文書の電子化は加速的に進みつつある。
【０００３】電子出版という言葉が示すように、今後は新聞、雑誌や本の情報も電子的に提供されることが一般的になると考えられる。これにより、個人にとってリアルタイムで入手可能となるテキスト情報の量は膨大になっていくと予測される。
【０００４】これに伴ない、これらの新聞や雑誌などの膨大なテキスト情報からユーザの要求・興味にあったものを選出してユーザに提供する情報フィルタリングシステムあるいは情報フィルタリングサービスの需要が高まりつつある。
【０００５】このような問題意識から、最近では、たとえばユーザごとに予め設定された検索条件に合致する情報のみをユーザに提供するといった情報フィルタリング装置が考慮され始められている。しかしながら、これらの情報フィルタリング装置にあっては、新聞記事や雑誌記事などといった定期的に発生する文書を処理対象とし

ているため、発生した文書の修正といった状況を考慮する必要がなかった。たとえば、新聞記事は毎日発行されるため、その日に発生した記事を対象にしてフィルタリング処理を行なえばよかった。また、ある程度の件数をまとめてＣＤ－ＲＯＭなどの媒体で定期的または不定期的に発行されるものについては、この発行されたＣＤ－ＲＯＭ内の情報のみを処理対象とすればよかった。

【０００６】このような状況は、作成された日付情報をその文書内に明示的に記述されている文書を対象とする場合も同様である。すなわち、作成された日付情報を参照して、その日付が所定の期間内にある文書だけを情報フィルタリングの処理対象とすればよく、容易に実現することができる。あるいは作成日や修正日などが、文書の補助情報として格納されるファイルシステムの場合でも、同様の処理が可能である。

【０００７】ところが、文書の中には、作成日や修正日などの日付情報が文書内に記述されておらず、また、文書ファイルの補助情報としても存在しておらず、さらに、その文書の作成の取り決めがされていないといった種類のものも存在する。たとえば、ＷＷＷ（Ｗｏｒｌｄ Ｗｉｄｅ Ｗｅｂ）で公開されている文書（Ｗｅｂページと呼ばれる）は、個人が何らコントロールされることなく作成することができる。そして、これらは個人の気の向いたときに作成され、しかも作成日や修正日などは文書内に記述するといった取り決めはない。このために、すべての文書について、その文書がいつ作成されたか、あるいはいつ修正されたかといったことを示す信頼性の高い日付情報をコンスタントに取り出すことは困難である。

【０００８】すなわち、従来の情報フィルタリング装置にあっては、新たに作成された（または修正が施された）情報の中から個人の興味にあった情報を選出して提供するという情報フィルタリング装置の目的に対して、これらの文書を特定することができないといった重大な問題があった。

【０００９】

【発明が解決しようとする課題】このように、従来の情報フィルタリング装置にあっては、その文書がいつ作成されたのか、またはいつ修正されたのかといった日付情報をすべての文書についてコンスタントに取得することが困難であるために、新たに作成された文書と修正が施された文書とを特定することができないといった問題があった。

【００１０】この発明は、このような実情に鑑みてなされたものであり、非定期的に作成および修正される文書であって、これらが生成または修正された時期を示す日付情報を持たない複数の文書の中から、新たに生成または修正された文書のみを検出してユーザに提示することを可能とする情報フィルタリング装置および情報フィルタリング方法を提供することを目的とする。

【００１１】

【課題を解決するための手段】この発明は、非定期的に作成および修正される複数の文書の中から所定の文書を選出してユーザに提示する情報フィルタリング装置において、新たに作成または修正された文書を検出する文書修正検出手段と、この文書修正検出手段によって検出された文書と予めユーザが指定した検索条件との間の類似度を算出する類似度算出手段と、この類似度算出手段により算出された類似度にしたがって前記検出された文書すべてを並べ換え、この並べ換えた順番でこれらの文書をユーザに送信または提示する手段とを具備してなることを特徴とする。

【００１２】また、この発明は、前記文書自体または前記文書を情報圧縮したデータを文書複製情報として各文書の文書ＩＤと対応させて格納する文書複製情報記憶手段をさらに具備し、前記文書修正検出手段は、前記文書複製情報記憶手段に格納された文書複製情報との比較によって前記文書の作成および修正を検出する手段を具備してなることを特徴とする。

【００１３】この発明においては、非定期的に作成および修正されるといった種類の文書であっても、これらの文書を定期的に取得し、かつその文書またはその文書を圧縮したデータを複製情報として記憶しておいて、取得した文書の情報と複製情報とを比較することによって、複数の文書の中から新たに作成または修正された文書のみを検出する。そして、この新たに作成された文書および修正された文書を情報フィルタリングの対象にすることにより、ユーザに対して新しい情報のみを提供することを可能とする。

【００１４】

【発明の実施の形態】以下、図面を参照してこの発明の実施の形態について説明する。まず、図１を参照してこの発明に係る情報フィルタリングシステムの全体の構成について説明する。

【００１５】図１に示すように、この発明の情報フィルタリングシステム１０は、文書データを格納する文書データベース１１、文書の複製情報を記憶する文書複製情報記憶部１５、文書データベース１１に格納された文書と文書複製情報記憶部１５に格納された文書との差分を求めることによって、文書データベース１１中に格納されたすべての文書の中から修正された文書のみを検出する文書修正検出部１２、ユーザの興味がある文書を検索するための検索条件を記述したプロファイルを格納したプロファイル記憶部１６、ユーザのプロファイルと文書との間の類似度を算出する類似度算出部１３、類似度算出部１３で算出した類似度にしたがって文書群をユーザに提示出力する文書提示部１４、および情報フィルタリングシステム１０全体の制御を行なう制御部１７からなる。

【００１６】ここで、この文書修正検出部１２、類似度

算出部１３、文書提示部１４および制御部１７の処理の流れをフローチャートに基づいて説明する。図２には、文書修正検出部１２の処理の流れが示されている。文書修正検出部１２は、制御部１０から定期的に起動されるものである。

【００１７】文書修正検出部１２では、文書データベース１１中のすべての文書を一つずつ取得しながら（ステップＡ１～ステップＡ３）、新しく作られた文書であるか否か（ステップＡ４）、すでに存在していた文書の場合には、その文書に修正が施されたか否か（ステップＡ６）を判定し、それらのいずれかの条件が成立したときに（ステップＡ４のＮ，ステップＡ６のＹ）、その文書を類似度算出部１３における類似度計算を行なう対象とする（ステップＡ５）。なお、ここで取得された文書すべては改めて文書複製記憶部１５に格納される（ステップＡ７）。そして、文書データベース１１中のすべての文書について繰り返し処理を行なうことにより、新規に作られた文書と修正された文書とを検出し、この文書をフィルタリング対象とする。

【００１８】図３には、類似度算出部１３の処理の流れが示されている。この類似度算出部１３は、文書修正検出部１２で検出されたすべての文書に対してプロファイル中の検索条件との間で類似度を算出する。検索条件と文書との間の類似度を算出するに当たっては様々な算出方法が開示されている。そして、類似度の算出方法に関しては、本発明では特に限定するものではなく、種々の方法を採用することが可能である。ここでは、検索条件と文書とをそれぞれ単語頻度のベクトルとして表現し、＊これらベクトル間の内積を取ることにより類似度を求める方法に基づいて説明する。

【００１９】すなわち、類似度算出部１３は、文書修正検出部１２によって検出された文書の文書ベクトルを算出し（ステップＢ２）、この算出した文書ベクトルと、プロファイル記憶部１６に格納されたプロファイルのベクトルとの内積を取って類似度を算出する（ステップＢ３）。

【００２０】図４には、プロファイル記憶部１６に格納される個人プロファイルの形式と一例とが示されている。図４中、（ａ）は形式を、（ｂ）は一例をそれぞれ示している。プロファイルは、単語と重みとのペアのリストからなる。（ｂ）の例では、「計算機」、「開発」、「新製品」といった単語の重みが２、１、１といったようにそれぞれ設定されている。

【００２１】図５には、文書側の内部表現例が示されている。この文書側の内部表現もプロファイルと同様に、単語と重みとのペアのリストからなっている。これは、文書の形態系解析を行なうことにより単語を抽出し、各単語の文書中の使用頻度を重みとして採用する。ただし、文書長の影響をなくすため、使用頻度をその文書中で最も多く用いられている単語の頻度、またはその文書中の全単語数などで割ることで正規化する。

【００２２】プロファイルか文書のいずれか一方に含まれる単語の重みを要素とするベクトルを考えた場合、図４の例の場合には、プロファイル側のベクトルｑは、次式で表現することができる。

【００２３】

計算機　開発　新製品　ＣＰＵ　メモリ　発売　今年

［２　　１　　１　　０　　０　　０　　０　］　…（１）式

一方、図５の例の場合には、文書側のベクトルｄｉは、次式で表現することができる。

※【００２４】

※

計算機　開発　新製品　ＣＰＵ　メモリ　発売　今年

［０．２　０．１　０　　０．１　０．１　０．２　０．１］　…（２）式

類似度計算として文書とプロファイルとの内積を取るものとすると、文書ｉのプロファイルに対する類似度Ｓｉは、次式で与えられることができる。

【００２５】Ｓｉ＝ｑ・ｄｉ　…（３）式

ただし、「・」は内積を意味している。そして、図４と図５の例の場合、Ｓｉは０．５となる。

【００２６】文書提示部１４では、各文書ごとに算出された類似度の大きい順に文書をソーティングしてユーザに提示する。一つの計算機に閉じた形態で本発明を実施した場合には、得られた文書群をソーティングされた順で直接ディスプレイなどに表示することになる。一方、ネットワークなどを介してユーザに送付する形態を取る場合には、ユーザには類似度順にソーティングされた文書群をファイル転送などの形で送ることになる。また、類似度が予め定められたしきい値を越えた文書のみをユーザに提供する文書としてソーティング対象としても構わない。

【００２７】制御部１７は、予め定め間隔で文書修正検出部１２を起動する。そして、この文書修正検出部１２によって新たに作成された文書または修正された文書が検出された場合に、類似度算出部１３を起動する。類似度算出部１３では、プロファイルとの類似度を算出する。さらに、１つ以上の文書の類似度が類似度算出部１３よって算出された場合に、文書提示部１４を起動する。文書提示部１４は、この算出された類似度にしたがって文書をソーティングしてユーザに提供する。

【００２８】すなわち、この情報フィルタリングシステムによれば、文書修正検出部１２によって、不定期的に作成および修正される文書を検出することができ、この検出された文書のみを対象にプロファイルとの間の類似度を算出してユーザに提供することが可能になる。

【００２９】（第１実施形態）ここで、この発明の第１

の実施形態について説明する。文書修正検出部12で格納する文書複製情報は、原文書そのままであっても構わないが、その場合には、原文書と同じ容量の記憶領域を必要とすることになってしまい、資源管理の面で好ましくない。したがって、文書複製情報として原文情報を圧縮して格納することにより、記憶容量の削減を図ることができる。

【0030】図6には、本実施形態における文書修正検出部12の処理の流れが示されている。前述の図2で示した文書修正検出部12の処理との相違は、取得した文書データに対してデータ圧縮を施す点にある（ステップC5，ステップC8）。データ圧縮の手法としては、すでに開示されている様々な手法をとることが可能である。たとえば、UNIXコマンドとして存在するcompressコマンドでは、適応型レンペル・ジブ・コーディング法が採用されている。このデータ圧縮の手法については本発明の主旨ではなく、あらゆる手法が採用可能である。

【0031】また、本発明を実施する場合、通常のデータ圧縮で保証されるデータの復元（圧縮する以前の元のデータに戻すこと）は、必ずしも必要ではない。これは、前回の文書情報取得時に同一の文書について算出したデータと比較して、相違があるか否かを判定するだけでよいためである。

【0032】たとえば、以下のような手法でデータ圧縮することも可能である。

```
for（i=0，a=0；buf[i]!=NULL；i++）
a^=buf[i]；
```

この例は、C言語での実現例であるが、配列buf中に文書iが格納された場合に、buf中のすべての文字について排他的論理和を算出している。この手法によれば、いくら長い文書であっても1バイトに圧縮されることになる。

【0033】（第2実施形態）次に、この発明の第2の実施形態について説明する。図7に、ネットワークを介して文書にアクセスする場合の情報フィルタリングシステムの構成を示す。前述の図1との相違は、文書データベース11がシステム内にはなく、ネットワーク20を介して文書にアクセスする点にある。本実施形態では、WebページをHTTP（HyperText TransferProtocol）によりアクセスする。Webページは、そのページの持ち主により不定期的に作成および修正される。そして、全体の処理フローは前述と同様であり、ネットワーク20と接続されて文書にアクセスするか、ローカルに文書データベース11と接続されているかの違いしかない。

【0034】なお、前述では、文書それぞれの日付情報を持たない文書データベースを適用した場合の例を示したが、作成日や修正日などを示す信頼性の高い日付情報がその文書と対応して記憶され、かつその日付情報が取り出せる文書データベースに対しては、文書複製情報を文書複製情報記憶部15などに記憶しなくとも、この作成日や修正日が前回のフィルタリング処理時点以降であったときに、文書が作成または修正されたものとすればよく、したがって、文書修正検出部12は、単に日付の比較を行なうだけで、新規作成および修正が判定できることになる。

【0035】また、この発明の手法は、ソフトウェアとして実現可能であるため、CD－ROMやフロッピィディスクなどといった記録媒体によって頒布することが可能である。また、磁気ディスクなどに格納しておき、ネットワークで取り寄せる（ダウンロード）ような形式で頒布することも可能である。

【0036】

【発明の効果】以上詳述したように、この発明によれば、不定期的に作成および修正される文書であって、かつこれらが生成または修正された時期を示す日付情報を持たない（またはその日付情報の信頼性が低い）種類の文書であっても、これらの文書を定期的に取得して、その文書またはその文書を圧縮したデータを複製情報として記憶しておき、取得した文書の情報と複製情報とを比較することによって、複数の文書の中から新たに作成または修正された文書の検出するため、新規に作成または修正された文書のみをユーザに送信することが可能となる。

【図面の簡単な説明】

【図1】この発明に係る情報フィルタリングシステムの全体構成図。

【図2】この発明の文書修正検出部の処理の流れを示すフローチャート。

【図3】この発明の類似度算出部の処理の流れを示すフローチャート。

【図4】この発明のプロファイル記憶部に格納される個人プロファイルの形式と一例とを示す図。

【図5】この発明に係る文書側の内部表現例を示す図。

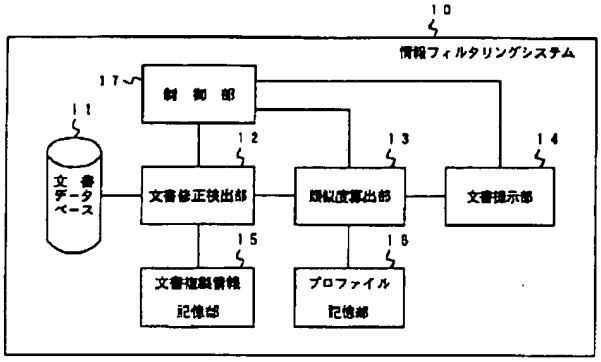
【図6】この発明の第1実施形態に係る文書修正検出部の処理の流れを示すフローチャート。

【図7】この発明の第2実施形態に係るネットワークを介して文書にアクセスする場合の情報フィルタリングシステムの構成を示す図。

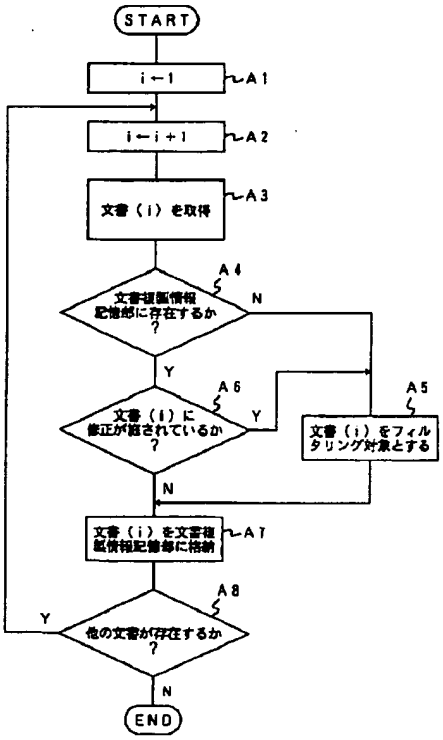
【符号の説明】

10…情報フィルタリングシステム、11…文書データベース、12…文書修正検出部、13…類似度算出部、14…文書呈示部、15…文書複製情報記憶部、16…プロファイル記憶部、17…制御部、18…ネットワーク接続部、20…ネットワーク。

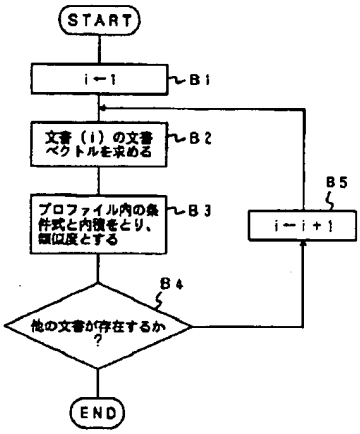
【図1】


10
情報フィルタリングシステム
17
制御部
11
文書データベース
12
文書修正検出部
13
類似度算出部
14
文書提示部
15
文書複製情報記憶部
16
プロファイル記憶部

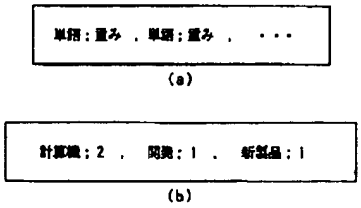
【図2】


START
i←1
A1
i←i+1
A2
文書（i）を取得
A3
A4
文書複製情報記憶部に存在するか？
N
Y
A6
文書（i）に修正が施されているか？
Y
A5
文書（i）をフィルタリング対象とする
N
文書（i）を文書複製情報記憶部に格納
A7
A8
他の文書が存在するか？
Y
N
END

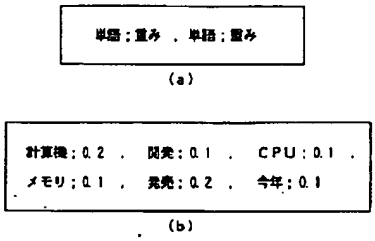
【図3】


START
i←1
B1
文書（i）の文書ベクトルを求める
B2
プロファイル内の条件式と内積をとり、類似度とする
B3
B4
他の文書が存在するか？
B5
i←i+1
END

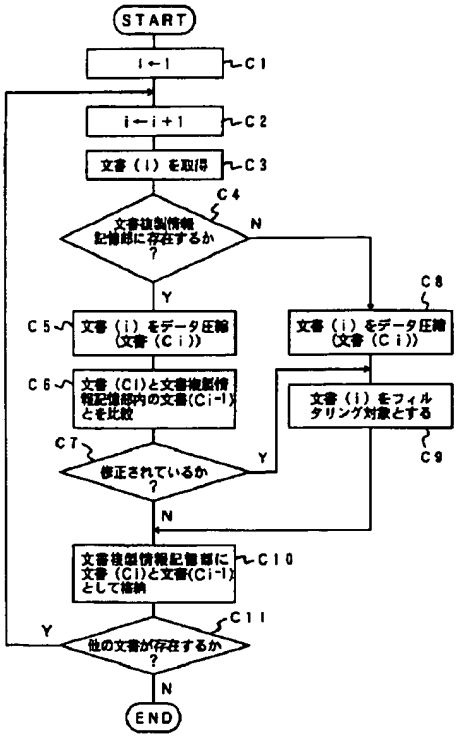
【図4】


単語：重み ，単語：重み ，・・・
(a)
計算機：2 ，開発：1 ，新製品：1
(b)

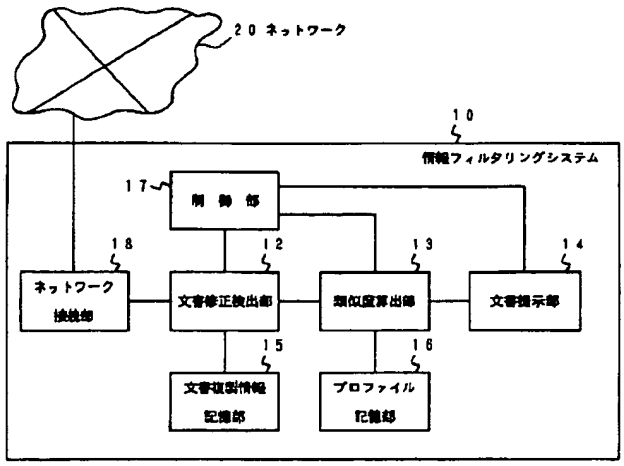
【図5】


単語：重み ，単語：重み
(a)
計算機：0.2 ，開発：0.1 ，CPU：0.1 ，
メモリ：0.1 ，発売：0.2 ，今年：0.1
(b)

【図６】



【図７】



フロントページの続き

(72)発明者　梶浦　正浩
神奈川県川崎市幸区小向東芝町１番地　株式会社東芝研究開発センター内